D209 Classification Analysis

Nhi Le

Western Governors University

September 4, 2023

**A1. Proposal of question**

Which customers are at high risk of churn? What are the features that significantly impact churn?

**A2. Defined goal**

From this analysis, the stakeholders can understand what factors affect churn to see where the services can be improved. Moreover, they will be able to predict which customers have high risks of retention. Then, the customer support team can contact those customers and have strategies to improve their experience with the company.

**B1. Explanation of classification method**

I will be using k-nearest neighbor (KNN) as the classification method for this analysis. This method uses the distance between the predicted point to all points in the data set to classify where the predicted point belongs to base on the majority label of the “k” closest points. K is the number of nearest neighbors. The expected outcomes are that a k number of data points is chosen and our data points in the test data set are classified with the K nearest neighbors.

**B2. Summary of method assumption**

One assumption of KNN method is that it is non-parametric, which does not make any underlying assumptions about the distribution of data (GeeksforGeeks, 2023).

**B3. Packages or libraries list**

Here are the packages and libraries I will use for Python:

* NumPy is used for working with arrays.
* Pandas is used for working with a data set. Example, we can use .read.csv() to load the data set to Python. Or I can use .info() to get the information of the data set.
* Matplotlib is a comprehensive library for creating visualizations. I will use matplotlib.pyplot submodule for creating histograms. They will help me detect outliers. This package also helps me to create bivariate visualization such as scatter plots.
* Seaborn is also used to create boxplots to detect outliers and create univariate/ bivariate visualizations.
* Scikit-learn is used for machine learning, especially for k-nearest neighbors. Example, I can use StandardScaler to perform feature scaling for independent variables. Train\_test\_split can be used to split the data set into training and testing sets.

**C1. Data processing**

My data pre-processing goal is to detect missing data, duplicate data, and outliers, then decide to treat them with appropriate methods. Moreover, I want to use all the independent variables for the analysis. To make that happen, I will need to encode categorical values into numerical values.

**C2. Data set variables**

The dependent variable used to answer the research question was ‘Churn’. It is a categorical variable as the values are ‘Yes’ or ‘No’.

The independent variables used to answer the research question:

* The 10 continuous variables including:
* ‘Children’, ‘Age’, ‘Income’ are demographic variables on billing statement for each customer.
* ‘Email’ is numeral variable to record the number of marketing or correspondence emails sent.
* ‘Contacts’ is numeral variable for how many times customer contacted technical support.
* ‘Outage\_sec\_perweek’ shows system outages in the customer’s neighborhood’s average of seconds per week.
* ‘Bandwidth\_GB\_Year’ (the average yearly amount of data used, in GB, per customer) is a continuous variable.
* ‘Tenure’ is numerical variable to record how many months the customer has been with the provider.
* ‘MonthlyCharge’ is the monthly charge for the customer.
* ‘Yearly\_equip\_failure’ is numeral variable to show the number of time customer’s equipment failed and needed to reset or replaced last year.
* The 23 categorical variables including:
* 8 categorical variables reflect customer’s satisfaction ratings on a scale of 1 to 8 (1 = most important, 8 = least important): ‘item1’ – Timely response, ‘item2’ – Timely fixes, ‘item3’ – Timely replacements, ‘item4’ – Reliability, ‘item5’ – Options, ‘item6’ – Respectful response, ‘item7’ – Courteous exchange, ‘Item8’ – Evidence of active listening.
* ‘Gender’ is categorical variable to reflect the gender of customer.
* ‘Techie’ has Yes/No value. This categorical variable reflects if the customer thinks that they are good at technology.
* ‘Contract’ is categorical variable on what kind of contract customer has ‘Month-to-month’, ‘One Year’, or ‘Two Year’.
* ‘Tablet’ is categorical variable answering if the customer has a tablet.
* ‘Port\_modem’ is categorical variable answering if the customer has a portable modem. The values are ‘Yes’ or ‘No’
* ‘InternetService’ is categorical variable that shows customer’s internet service provider.
* ‘Phone’, ‘Multiple’, ‘OnlineSecurity’, ‘OnlineBackup’, ‘DeviceProtection’, ‘TechSupport’, ‘StreamingTV’, ‘StreamingMovies’, ‘PaperlessBilling’ are services that the company provides. The values of these variables are ‘Yes’ or ‘No’ to reflect if the customer signed up for. These are categorical variables.

**C1. Steps of analysis**

Data preparation steps:

* Import dataset churn\_clean.csv into Jupyter Notebook.
* Get information (column names, data types), and statistical details (count, min, max, mean, std, percentile) of the dataset.
* Detect duplicates and delete the duplicated records if there are any.
* Find missing data and impute missing data with meaningful measures of central tendency (mean, median, or mode).
* Find outliers and treat them by removing them, retaining them, excluding them, or imputing them with the median.
* Run univariate and bivariate visualizations to see the spread of data.
* Rename the Item1 – Item8 columns to easily recognized names (For example: ‘Item1’ renamed to ‘TimelyResponse’).
* Drop variables that will not be needed for the analysis.
* Create dummy variables for categorical variables.
* Encode categorical values to numerical values: For those variables with Yes/No values, the dummy value is 1 for Yes and 0 for No. For the Gender variable, it has Male, Female, and Nonbinary. The DummyFemale is 1 when Gender is Female and else it is 0. Contract has 3 values: Month-to-month, One Year, and Two year. DummyMonthtoMonth is 1 when Contract is Month-to-month, else it is 0. InternetService has 3 values: Fiber Optic, DSL, and None. DummyFiberOptic is 1 when InternetService is Fiber Optic, else it is 0.
* Spot-check the statistical details of the dataset to make sure categorical values are encoded correctly.
* Drop those categorical values from the data set.
* Extract the prepared dataset as CSV file named ‘churn\_prepared.csv’.

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A white rectangular object with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

A line of dots on a white background

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A graph with a bar and a line

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screen shot of a computer

Description automatically generated

A screen shot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**C4. Cleaned data set**

The prepared data set will be submitted as ‘churn\_prepared.csv’ along with this doc file.

**D1. Splitting the data**

The csv files will be submitted as ‘X\_train.csv’, ‘X\_test.csv’, y\_train.csv’, and ‘y\_test.csv’ along with this doc file. The data set was split into training (80%) and testing (20%) sets.

A screenshot of a computer program

Description automatically generated

**D2. Output and intermediate calculations**

Once I created my training and test data sets, I normalized the independent features. The data normalization should happen after splitting the data set into training and test sets. This would prevent ‘data leakage’ as the normalization would give the model additional information about the test set if we normalized all the data at once (Shafi, 2023). I then fitted the data sets into the model and created a new array called ‘pred’ to make prediction on the test data set. The analysis technique I used to appropriately analyze the data is by calculating the accuracy score of the KNN model. I used Cross Validation to get the best value of k. As the k value ran from 1 to 40, I could pick the k value with the highest accuracy score, which was 25.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**D3. Code execution**

The code was included in D2 above, and in ‘D209(1) NL.ipynb’ file submitted along with this doc file.

**E1. Accuracy and AUC**

The accuracy score of the KNN model with K=25 is 0.86, out of 2000: True Positive + True Negative = 1389 + 328 = 1717, False Positive + False Negative = 230 + 53 = 283. The AUC (Area Under Curve) of the model with K=25 is 0.92, which is an acceptable score.

A screenshot of a computer

Description automatically generated

**E2. Results and implications**

The accuracy score of the KNN model with K=1 is 0.79. By using cross validation, I could choose the best K, K=25, the accuracy score improves to 0.86. The AUC scores improved from 0.73 to 0.92 for model with K=1 and K=25. Therefore, the KNN model with K=25 is a better model. In the analysis, I used 10 continuous and 23 categorical variables to predict churn and to classify whether customers were at high risk of churn. Even after normalizing the independent variables and implementing cross validation to pick the best K, the accuracy score of the model is still relatively low. In conclusion, I do not think the model is significantly practical.

**E3. Limitation**

The limitation of the data analysis is that the KNN has slow speed. When I ran the cross validation, it took a long time to complete. Moreover, the more the number of features is used, the worse KNN's performance becomes because of the curse of dimensionality. Therefore, the results of my analysis were not significantly practical since I used a lot of features in the model.

**E4. Course of action**

Even though the accuracy score of the model is at 0.86, the stakeholders can still use this as a reference on top of other regressions, such as logistic model. The stakeholders can classify the group of customers who were at high risk of churn. Then, the recommended course of action is to analyze the features that appear in common among this group. For example, the stakeholders can look at whether high cost of services or low bandwidth limit are the reasons for churn. From there, they can have a better pricing strategy to make it more affordable for users. Also, they can investigate the data demand for customers and have appropriate data plans for customers to give them smooth services.

**F. Panopto recording**

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=bdb0b9d6-d733-4ad5-8743-b073017cc139>

**G. Sources for third-party code**

Shafi, A. (2023, February 20). *K-Nearest Neighbors (KNN) classification with scikit-learn*. DataCamp. https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn

**H. Sources**

GeeksforGeeks. (2023c, May 5). *K-Nearest Neighbor(KNN) algorithm*. GeeksforGeeks. https://www.geeksforgeeks.org/k-nearest-neighbours/

Shafi, A. (2023, February 20). *K-Nearest Neighbors (KNN) classification with scikit-learn*. DataCamp. https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn